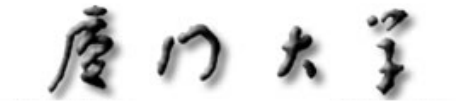
学校编码：10384　　分类号＿＿＿＿密级＿＿＿＿
学号：15420111151892　　UDC＿＿＿＿＿＿＿＿



硕 士 学 位 论 文

基于统计学的个性化推荐算法探究

Research on Personalized Recommender Algorithm Based on Statistics

黄秋婷

指导教师姓名：周永强　副教授

专　业 名 称：统 计 学

论文提交日期：2014 年 3 月 20 日

论文答辩时间：

学位授予日期：

答辩委员会主席：＿＿＿＿＿＿

评　　阅　　人：＿＿＿＿＿＿

2014 年 3 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

2014 年　月　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（ √ ）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

2014 年 3 月 20 日

摘　要

随着互联网的高速发展，信息呈爆炸式地增长，大数据在飞速的发展中，数据挖掘是一个充满活力的研究领域，商业利益的强大驱动力将会不断地促进它的发展，个性化推荐就属于大数据时代数据挖掘应用在互联网方面的重要技术。面对海量数据，推荐系统的产生能实现信息消费者和生产者的双赢。协同过滤算法是个性化推荐中最成功和应用最广泛的算法之一，但它依赖于用户的历史评分数据，所以存在冷启动，数据的稀疏性等问题。

大数据新形势下，包括个性化推荐在内的各种数据挖掘算法给统计学带来了机遇和挑战，一方面，数据挖掘的各种算法很多思想都来自于统计学；另一方面，数据挖掘面对统计学表现出了强大的生机。据此，本文探究在数据量比较大的情况下，将统计分析应用到个性化推荐算法中的效果，同时也应用数据挖掘的其他模型，如关联法则，聚类等方法改进模型。

本文提出了基于统计学的个性化推荐，主要是利用 MATLAB，SAS 进行辅助编程，分别实现了描述性统计、多维关联法则、协同过滤的算法进行推荐。对协同过滤模型存在的缺点的改进，针对模型的数据稀缺性和冷启动问题，结合用户的评分和特征信息，提出用一维和二维的统计量改进数据的稀疏度问题，然后利用 SQL SERVER 2005 和 EXCEL 数据挖掘外接模块对用户建立聚类模型，基于各类的统计分析改进模型，聚类模型不仅能解决数据的稀缺性，而且能克服冷启动问题；最后通过奇异值分解方法改进算法，并由平均绝对误差来衡量各种改进效果。通过对比本文得出结论：根据用户的评分和特征进行统计分析，用分析结果改进协同过滤算法有比较好的效果，结合统计学，数据挖掘的模型对于冷启动问题有较大的改善。本文的实验可以说明统计学的思想在各种复杂的模型中都能得到体现，在未来大数据发展的路上，统计学既要保持其最基础的生命力，同时要加强在其他学科的应用，推进统计方法制度改革，扩展统计学研究具体科学的深度和广度。

关键词：统计学；大数据；协同过滤；奇异值分析；聚类；个性化推荐

Abstract

With the rapid development of the Internet, information has exploded in growth. Data mining has became a vibrant researching area as the big data develops. Powerful driving of commercial interests will force it to continue to promote its development. Personalized recommendation is one of important technology which data mining is used on the Internet. Faced with massive data, recommendation system can achieve a win-win for information consumers and producers. Collaborative filtering algorithm is one of the most successful and the most widely used algorithms among personalize recommendations. However, it depends on the history data of users, so there are some problems such as cold start, data sparsity and so on.

Under the new situation of big data, all data mining algorithms, including personalized recommendations, bring the opportunities and challenges to Statistics. On the one hand, data mining algorithms have many ideas from the Statistics, on the other hand, data mining shows a strong vitality. Accordingly, the article is wrote to explore the effect when statistical analysis applied to personalized recommendation algorithm in the case of large amount of data and other data mining models such as the association rules, clustering and other methods to improve the model .

This paper presents a personalized recommendation based on statistics, mainly using MATLAB, SAS to do auxiliary programming, and bring out descriptive statistics, multidimensional association rules and collaborative filtering algorithm to recommend.

To improve existing shortcomings of collaborative filtering model, this paper present to use one-dimensional and two-dimensional statistics to improve the data sparsity problem, then use SQL SERVER 2005 and EXCEL data mining module to set up a clustering model. Finally, singular value decomposition method is used to improve the algorithm, and use average absolute error to measure the effect of various improvements. By comparing the conclusions we find that using the results of statistical analysis based on user ratings and characteristics to improve collaborative filtering have better effects. Combining statistics, data mining models for cold-start problems are greatly improved. Experiments described in this paper can be thought that statistical theory can be showed in a variety of complex models. In the way of the future development of big data, statistics not only maintain the most basic vitality, but also strengthen its application in other disciplines, promote statistical methods

reform, expanse the depth and breadth of statistical studies of specific scientifical studies.

Keywords: Statistics; Big Data; Collaborative Filtering; SVD; Clustering; Personalized Recommendation

目　录

Contents

第一章 绪论

1.1 课题的背景和意义

1.1.1 选题背景

随着互联网的发展，人们正处于一个信息过载的时代。近年来，对大数据的研究和应用不仅引起了我国自然科学和人文社会科学界的广泛重视，也受到我国中央政府的高度关注：《“十二五”国家战略性新兴产业发展规划》明确提出支持海量数据存储、处理技术的研发与产业化。大数据在社会生活中的应用，横跨金融交易、电子商务、决策制定、广告营销、医疗用药等。

对于统计理论的发展而言，大数据时代带来的不仅是变革，更是统计学发展壮大的机会。大数据将改变传统统计学研究具体问题的方法科学，改变统计研究的工作程序，改变统计学研究具体科学的深度和广度。2012 年 11 月国家统计局总统计师鲜祖德在会见美国华裔大数据专家学者时，明确提出国家统计局十分重视大数据在统计中的应用，并成立了专门的课题组着手研究如何通过对大数据的处理推进统计方法制度改革，改进政府统计工作。2012 年 12 月国家统计局在上海开展了大数据应用的调研活动。2013 年 2 月国家统计局召开了以大数据为主题的工作会议，3 月 4 日国家统计局科研所重点讨论了大数据的研究，同时部署了“大数据在政府统计中的应用”的研究工作，3 月 26 日科研所又举办了“大数据在政府统计工作中的应用研究”课题研究专家咨询会。4 月 11 日，国家统计局大数据课题组赴百度公司调研，10 月 28 日-29 日，“第十七次全国统计科学讨论会”在浙江省杭州市召开，其主题是大数据背景下的统计。虽然我国大数据的理论研究和应用研究刚刚起步，但是学术界、企业界及政府部门对该领域的重视程度前所未有。

IBM 及牛津大学 2012 年 10 月发布问卷调查报告称，大数据的主要分析能力有：1.查询与报 91%；2.数据挖掘 77%；3.数据可视 71%；4.可预测的建模 67%；5.优化分析 65%；6.模拟分析 56%；7.自然语言文本分析 52%；8.地理空间分析 43%；9.数据流分析 35%；10.视频分析 26%；11.声音分析 25%。可以看出数据挖掘是大数据时代一个非常重要的应用。

大数据的应用中，数据挖掘的核心技术包括了统计学、数据仓库和机器学习技术，统计学渗透于各种技术之中。这其中推荐系统是一个非常热的研究点，国内知名电子商务网站(淘宝，京东等)、社交网站(如豆瓣)都成功引入推荐系统，进行商品推荐，用户推荐等。互联网用户个人没有足够经验，无法从过载的信息量中快速准确地找到自己想要的信息，为此推荐系统出现，并解决这一难题，随后被成功引用到电子商务中，展示了强大的商机。推荐系统不仅限于导购，在互联网的高速发展中，推荐系统以后会更加广泛地应用于各个领域，如大学的图书馆系统，互联网交友平台，甚至是ＳＮＳ好友推荐等系统中。但是传统个性化推荐方法还存在很多问题，比如冷启动，数据的稀疏性等问题。目前应用最广泛的协同过滤推荐算法依赖于历史评分数据，对用户特征和项目特征信息不能充分加以利用，因此准确性很低，不能有效地向用户进行推荐，为了得到更好的推荐效果，越来越多的学者开始致力于研究高准确度和效率的推荐算法。

1.1.2 选题意义

个性化推荐系统目前主要是基于互联网平台，利用电子商务平台提供的用户信息，结合项目产品信息，分析用户的行为习惯，给用户产生推荐。这不仅大大节省了用户浏览信息的时间，也能让用户更快捷地找到自己满意的商品。对于商家来说，个性化推荐系统不仅可以用做企业的营销手段，更重要的是可以增加用户粘性。面对剧烈的市场竞争和飞速发展的电子商务，客户有更多的选择，信息的流通更加畅通，很多用户的粘性比较低。所以，去除价格因素后，如果个性化推荐系统的精确度较高的话，可以在很大程度上留住用户，为企业带来巨大的商业利益。

在此背景下，面对大数据时代给统计学的挑战和机遇，本文提出从统计学的视角进行推荐并改进个性化推荐系统，探究在数据量比较大的情况下，统计分析应用到数据挖掘，个性化推荐算法中的效果，同时也应用数据挖掘的其他模型，如关联法则，聚类等方法改进模型。这两种模型是现在数据挖掘中比较重要和常用的方法，在现代互联网的发展中对海量数据的处理，分析，预测，对企业提供决策支持等方面取得了非常重要的作用。这些方法中也传递着统计学的思想，对比其他更加智能的模型和方法，统计学是在与其他学科紧密结合应用中形成的学

科，统计学的生命力在于应用，在海量数据的时代更应该加强同其他学科的交叉协作。

本文主要采用个性化推荐最著名的网站 movielens 上的数据进行一个仿真研究，希望对于推荐系统的算法有个更加深入的理解，同时利用基本的统计方法提出一些解决方法，进行模拟仿真，以期通过这些简单的统计方法来思考在海量数据的分析中，传统的统计方法在复杂模型和算法中的作用。从统计学的角度去改进个性化推荐系统，是在大数据时代对于统计学应用的一种思考和探究。通过结合数据挖掘的模型，将统计学应用于个性化推荐，尝试让统计学在大数据时代能够适应数据量的变化，继续发挥其处理数据、分析数据的重要作用，期望自己能对统计学理论在大数据应用方向的发展做出思考和探究。

1.2 文献综述

1.2.1 数据挖掘与统计学的研究

国外对于统计学和数据挖掘方面的研究开始的比较早，而国内统计学界也开始慢慢关注着统计学与大数据的发展，中国人民大学统计学院是国内较早开展数据挖掘应用和理论探索的团队，是在 2001 年春季成立了数据挖掘研究中心，并且开辟了“统计学与数据挖掘”研究专栏，是比较早在统计学科下研究数据挖掘的团队。除此之外，厦门大学数据挖掘研究中心（Data Mining Research Center，DMRC）组建于 2006 年，由辅仁大学谢邦昌教授、厦门大学统计系的朱建平教授和美国耶鲁大学马双鸽副教授联袂领衔。中心充分发挥学科交叉优势，是在整合统计学、计量经济学、数量经济学、信息管理学等相关学科基础上成立的跨学科、综合型、开放式的教学研究机构。谢邦昌教授致力于将统计学应用发展，促进数据挖掘的研究和发展，有很多的研究成果。此外，包括东北财经大学、西南财经大学、国家统计局在内的，越来越多的机构和学者开始从关注数据挖掘，不断扩展统计学的广度和深度。

厦门大学统计系的朱建平（2005）教授是国内在统计学科进行数据挖掘研究比较活跃的学者，他研究了数据挖掘中的统计学方法，并且出版了专著《数据挖掘的统计方法与实践》；谢邦昌(2004)教授将数据挖掘的方法应用于信用卡、保险、金融投资等方面，对数据挖掘中关联规则作了统计描述，并深入研究了相应分析的适应性问题。王政霞、黄大荣（2005）在统计方法的基础上提出了一种根据数

据集合本身的统计特性数据挖掘算法。胡军刚(2008) 从统计学的角度对数据挖掘进行应用性研究，采用了描述性统计推荐和预测性统计推荐，根据聚类分析的基本思想提出了改进后的动态聚类分析方法，体现了定性分析(描述性挖掘部分)与定量分析(预测性数据挖掘部分)的结合；

1.2.2 个性化推荐系统的研究

个性化推荐系统是一种高级商务智能平台，它建立在海量数据挖掘基础上，给电子商务网站的顾客购物提供个性化服务，帮助决策支持和信息服务。协同过滤是个性化推荐系统最主要的算法，这一概念最早应用于 Tapestry 系统，是 1992 年由 Goldberg、Nicols、Oki 及 Terry 提出，主要是为了解决 Xerox 公司在 Palo Alto 的研究中心资讯过载的问题。

在早期的协同过滤技术中，系统只有在用户了解彼此间的兴趣爱好之后才能做出推荐；之后随着研究的不断深入，出现了自动化的协同过滤系统，这其中典型的代表就是由美国明尼苏达州立大学 GroupLens 项目组研发的 GroupLens 系统。GroupLens 系统首先建立用户群，群内用户可以发布自己感兴趣的信息，系统通过信息过滤系统度量用户之间兴趣的相似性，利用相似用户的信息进而向目标用户进行协同推荐。

随着电子商务的推荐，个性化推荐算法已经取得了很大的进展，有很多的研究成果，广泛应用于在电子商务推荐领域。主要的推荐方法包括：基于人口统计学的推荐、基于关联规则的推荐、基于内容的推荐，协同过滤推荐和混合推荐。但是，在海量数据下，随着用户和资源数目的不断增加，传统的协同过滤技术面临着巨大的挑战，主要表现在数据的高维稀疏性、算法的可扩展性和实时性方面。研究者从各个方面来降低数据稀疏带来的问题。例如 Sarwar 等人（2000）利用奇异值分解（SVD）的方法来减少用户-项目评分矩阵的维度，降维后得到相对稠密的数据，可以较好的解决数据稀疏问题，但推荐精度有所降低。此外，Goldberg（2001）用主成分分析对协同过滤进行降维。Karypis 等人（2001）提出了基于项目的协同过滤算法（Item-based CF），由于基于项目的相似性比基于用户的相似性稳定，在一定程度上缓解了数据稀疏问题。

同时越来越多的学者从数据挖掘，人工智能，数学等方面对模型进行了改进，通过采用聚类、贝叶斯网络、神经网络等手段来降低数据稀疏性。Min 等人（2005）

提出模糊聚类，对解决数据稀疏性带来的冷启动问题有很好的效果，而且通过离线计算不会给推荐系统的实时性带来负担。唐灿(2006)介绍了一种基于模糊兴趣模型的个性化推荐算法。马宏伟等人（2009）总结了协同过滤推荐算法中的关键问题和相关解决方案，提出利用 BP 神经网络技术来解决数据稀疏性问题。程中伟（2011）采用关联法则的算法进行个性化推荐，分析商品之间以及商品与用户之间的关系，并依此进行基于关联规则的数据挖掘。夏培勇（2011）提出了一种基于用户间评分差异信息熵的相似性度量方法（简称为 NWDE 算法），他还提出了一种改进的自适应 AdaBoot.RT 集成学习算法改进系统。赵丽嫚（2013）提出采用遗传算法，将用户的评分、特征因素与项目的评分、特征因素结合起来，搜索最佳参数组合，提高了推荐算法执行的效率。

1.3 论文研究内容及结构

1.3.1 研究内容

本文在收集和整理了大量的国内外相关资料后，在对个性化推荐系统的理论和方法有了一定的理解和体会的基础上，从统计学的视角提出自己对于个性化推荐系统的想法，主要围绕以下几个方向进行展开：

第一，介绍个性化推荐的相关理论，结合数据挖掘的算法和模型，详细地介绍了个性化推荐的各种算法及具体的实现思路。其中，以关联法则和协同过滤算法为主要内容，同时对于各种算法的优缺点有一个比较详细的整理，分析了推荐系统的评价方法和改进思路，为后面的模型实现做好理论基础.

第二，理解选题数据，对数据进行相应的清洗工作。分析整理选题数据的特点，进行预处理，并对数据进行划分和检验，选用不同的算法对数据进行处理和分析。

第三，基于统计学分析的个性化推荐，对数据进行基本的处理，从一维，二维的角度对数据进行统计分析，实现基于内容的个性化推荐，研究从不同的属性分类的角度去统计数据，进行简单的内容推荐。

第四，对协同过滤模型存在的缺点的改进，针对模型的数据稀缺性和冷启动问题，结合用户的评分和特征信息，提出从一维和二维的角度改进数据的稀疏度问题，然后对用户建立聚类模型，基于各类的统计分析改进数据稀疏性和冷启动

问题，最后通过奇异值分解方法改进协同过滤算法，并由平均绝对误差来衡量各种推荐效果。

1.3.2 论文的结构

本论文主要研究思路和结构是：：

第一部分是绪论部分。主要介绍了论文的研究背景、研究意义、统计学与个性化推荐系统研究的一些基本的内容和综述，同时也对本文的研究内容和结构安排有了一个比较明确的整理。

第二部分是理论综述部分。本文从大数据时代，统计学和数据挖掘的各种算法和发展展开，介绍了基于人口统计学的推荐，基于内容的推荐，基于规则的推荐，协同过滤算法，并评价了各种算法的优缺点，分别介绍了协同过滤算法的几种方法及算法流程，整理了协同过滤算法的评价和改进。

第三部分是数据整理和统计分析。主要是介绍本研究所基于的数据来源和数据的说明，在对理解数据的基础上，从一元，二元不同维度对数据进行分析，整理结果分析后进行简单的推荐，同时分析得到的数据也便于我们对后面的模型和仿真实验。

第四部分是基于规则的推荐。本文对所采用的数据进行整理后，通过关联法则发掘推荐规则，并对结果有一个解释和分析，这里采用与 SQL SEVER 链接的 EXCEL 的数据挖掘功能实现多维的关联法则。

第五部分是协同过滤算法的实现。首先对数据进行划分并验证后，根据协同过滤算法的流程用训练集里的数据进行仿真计算，这里主要是采用 MATLAB 来实现仿真过程，并对于实验的不同参数进行计算，得到实验的评价指标，进行对比分析。

第六部分是协同过滤算法的改进。针对之前分析中算法存在的数据稀缺性和冷启动问题，本文提出从统计分析出发，分别从一维和二维不同角度对模型进行改进，同时结合奇异值分解方法和用户聚类的方法，通过 MATLAB 和 EXCEL 数据挖掘中的聚类模块来改进模型的数据稀缺性和冷启动问题，提高模型的推荐效果，最后同样计算评价指标，与前面的实验结果进行对比分析。

第七部分是结论及展望。主要是提出本文的研究总结，并提出相关建议和以后的发展方向。